When temporal expressions help to detect vital documents related to an entity

Rafik Abbes, Karen Pinel-Sauvagnat, Nathalie Jane Hernandez, Mohand Boughanem

When Temporal Expressions Help to Detect Vital Documents Related to An Entity

Rafik Abbes
IRIT, University of Toulouse
rafik.abbes@irit.fr

Karen Pinel-Sauvagnat
IRIT, University of Toulouse
karen.sauvagnat@irit.fr

Nathalie Hernandez
IRIT, University of Toulouse
nathalie.hernandez@irit.fr

Mohand Boughanem
IRIT, University of Toulouse
mohand.boughanem@irit.fr

ABSTRACT

In this paper we aim at filtering documents containing timely relevant information about an entity (e.g., a person, a place, an organization) from a document stream. These documents that we call vital documents provide relevant and fresh information about the entity. The approach we propose leverages the temporal information reflected by the temporal expressions in the document in order to infer its vitality. Experiments carried out on the 2013 TREC Knowledge Base Acceleration (KBA) collection show the effectiveness of our approach compared to state-of-the-art ones.

CCS Concepts

•Information systems → Information retrieval; Document filtering;

Keywords

Entity, vital documents, temporal expressions, document filtering, TREC Knowledge Base Acceleration

1. INTRODUCTION

Knowledge bases such as Wikipedia and Freebase are among the main sources visited by users to access knowledge on a wide variety of entities [12]. Accessing this information might seem easy, but it must be tempered by the freshness of information that can be found in the knowledge bases. With the growing amount of information available on the web, it becomes more and more difficult to detect relevant and new information that can be used to update knowledge base entities. Frank et al. [10] showed that the median delay of update can reach 365 days for wikipedia articles related to a sample of non-popular entities, which makes many knowledge bases entries out of date. This gap could be reduced if timely-relevant information could be automatically detected as soon as it is published and then recommended to the editors.

Let t_0 be a reference date corresponding to the date of the last update of the knowledge base entity (e.g., the Wikipedia page of the entity). A system that analyses a stream of documents to filter those providing new entity related information that was not known at t_0 can be very useful for knowledge base editors. We call these documents *vital* documents.

The most challenging aspect is to draw a distinction between *old-relevant* documents (also called *useful*) and *vital* ones. The former provide entity related information that was known at t_0, whereas the latter reveal new relevant information about the entity that was not known at t_0 and they are more likely to be helpful to maintain an already up-to-date knowledge page entity.

In Figure 1, we consider the entity *Michael Schumacher* and the reference date t_0= *December* 2013. Document D does not provide any information about the entity, it is therefore *non-relevant*. Document A is *old-relevant* as it reports only old relevant information that is known at t_0. Document C contains timely relevant information about *Michael Schumacher*, and is thus considered as *vital*. Obviously, a previously *vital* document for the entity will become *old-relevant* in the future, for example document B is vital if t_0 = *December* 2013, but probably not after a couple of months.

Most of the state-of-the-art approaches [2, 7, 8, 15] focused on detecting documents that are relevant for the entity without distinguishing between *old-relevant* and *vital* ones. In this paper, we attempt to make this distinction and we aim at detecting only vital documents for the entity.

The approach we propose is based on leveraging the delay between the *publication date* of the document and the dates referred to by the *temporal expressions* (dates, days, times, etc.) that can exist in the document text. We think that when this delay is short, the document matching the entity is more likely to be *vital*. For example, in Figure 1, by analysing the temporal expressions mentioned in the text of the documents A and B, we can guess that the second document is more likely to be vital for the entity as it mentions temporal expressions (*December 31,2013*, *Tuesday*) that refer to a date close to the publication date of the document, whereas document A contains old dates (*2006*, *2012*) compared to the publication date (*2014*).

To the best of our knowledge, this is the first work that uses temporal expressions expressed in the document to infer its

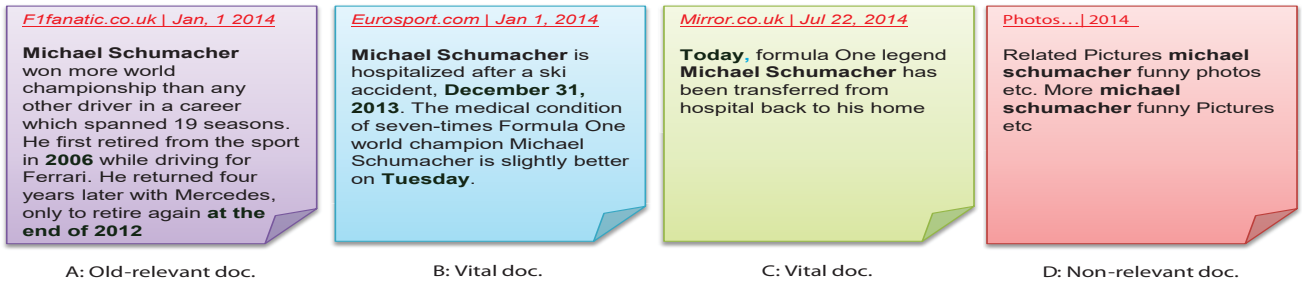

Figure 1: Distinction between *non-relevant*, *old-relevant* and *vital* documents assuming that the reference date t_0= *December* 2013

vitality for a given entity. Experiments carried out on the 2013 TREC KBA collection show the effectiveness of our approach against state-of-the-art ones.

The remaining of the paper is organized as follows. Section 2 reviews some related work. Section 3 describes the approach we propose. In Section 4, we report and discuss the experimental results carried out on the 2013 CCR collection of the TREC KBA track. We conclude and suggest some future work in Section 5.

2. RELATED WORK

In recent years, more and more attention has focused on filtering entity-centric documents. For instance, the TREC *Knowledge Base Acceleration* (KBA) has investigated the challenge of detecting relevant documents about a specific entity (e.g., a person, an organization, a place) since 2012. Particularly, the *Cumulative Citation Recommendation* (CCR) task strives to recommend to the editors of a knowledge base, relevant documents from an incoming document stream [9, 10, 11]. Since 2013, the CCR task asked participants to distinguish between *old-relevant* documents (called *useful* in the terminology of TREC KBA) and *vital* documents (containing timely relevant information about the entity). The proposed approaches can be classified into three categories.

The first category of approaches tackled the task as *a ranking problem* [7, 13]. Dietz and Dalton [7] used query expansion with related entity names to retrieve relevant documents. Liu et al. [13] ranked documents that match the entity by leveraging the number of occurrences and weights of related entities collected by parsing the Wikipedia page of the entity. The previously described works perform well when filtering relevant documents for the entity. However they do not attempt to distinguish *vital* documents from *old-relevant* ones. One possible reason is that these relevance models attempt to capture topicality rather than temporal characteristics.

The second category tackled the task as a *classification problem* [3, 4, 14]. Bonnefoy et al. [4] as well as Balog et al. [3] proposed one-step and multi-step classification approaches that attempt to learn the relevance of a document based on four families of features: (1) *document features* such as document length and document source; (2) *entity features* such as the number of related entities from DBpedia; (3) *document-entity features* that describe the relation between a document and the entity such as the number of occurrences of the entity in the document; and (4) *temporal features* which attempt to capture if something is happening around the entity at a given point in time by analysing the changes in the stream volume and in the number of views in the Wikipedia page of the entity. Wang et al. [14] used the Random Forest classifier trained on human-annotated documents. The same four families of features as in [3] were used, with in addition, *citation features* reflecting the similarity between a new document and cited documents in the Wikipedia page of the entity.

The previously reported works leverage *temporal* features that attempt to capture the entity related "bursts" (changes in the entity Wikipedia page views, or in the stream). These features have been shown to perform well [3]. In this work, we use another kind of temporal evidence to detect vital documents. We exploit *temporal expressions* in the document which, we believe, have not been previously used for this particular problem.

The first two categories of approaches assign a score for each document based on the probability of it belonging to a relevant class, or on a scoring function. Efron et al. [8] proposed a third kind of approach based on learning boolean queries that can be applied deterministically to filter relevant documents for the entity. In this work, we apply some heuristic boolean filtering rules that help to reject many non-vital documents (Section 4).

3. LEVERAGING TEMPORAL EXPRESSIONS FOR FILTERING VITAL DOCUMENTS

This paper is concerned with the task of filtering vital documents related to an entity: Given an entity E and a reference date t_0, we aim at identifying from a stream of documents those that are vital to the entity (i.e. containing new relevant information not known at t_0).

Generally, tackling this task involves two main steps: *filtering* then *scoring*. The filtering step can be seen as a way to eliminate many non-relevant documents (for example documents not mentioning the target entity). We also used this step and we detail some filtering rules in Section 4. In this section, we focus on the scoring step and we assume having a set of candidate documents that were filtered.

Our idea is to consider that a vital document related to an entity should report information about this entity and also should be fresh. *Freshness* can be determined by checking the publication date of the document and the temporal expressions used in its text. We assume that a vital document should be recent (published after the reference date t_0), and report a date greater than t_0 and close to its publication date.

In Figure 1, documents A and B are topically relevant to the entity *Michael Schumacher* and both were published on *Jan 1, 2014*. Document B is fresher than A because it mentions temporal expressions that refer to a date close (*Tuesday December 31,2013*) to the publication date of the document, whereas document A contains old dates (*2006*, *end of 2012*) compared to the publication date (*2014*).

Given an entity E and a new candidate document d published on $Date_p(d)$. Let $Date_t^*$ be the closest date (to $Date_p(d)$) recognized from the part of text mentioning the entity E in d. We assume that the shorter the period between $Date_p(d)$ and $Date_t^*$, the higher the probability of document d to be vital for E will be. Formally, we evaluate a freshness score of the document d with regard to the entity E as follows:

$$Freshness(d, E) = e^{-\frac{1}{\sigma^2}\Delta(d,E)} \quad (1)$$

$$\Delta(d, E) = \min_{x \in X(d,E)} (|Date_p(d) - Date_t(x, d)|^2) \quad (2)$$

Where

- $Date_p(d)$ is the publication date of d.
- $X(d, E)$ is the set of temporal expressions detected from the parts of d (sentences, paragraphs, etc.) that mention entity E.
- $Date_t(x, d)$ is the date indicated by the expression x.
- $\Delta(d, E)$ is the optimal (minimum) delay between $Date_p(d)$ and $Date_t(x, d)$ $\forall x \in X(d, E)$.
- $Date_t^*$ corresponds to $Date_t(x, d)$ where $\Delta(d, E)$ is minimal. $Date_t^*$ should be greater than the reference date t_0, otherwise the document is rejected as it is more likely to be *non-vital*.
- The maximum value of $Freshness(d, E)$ is equal 1 when document d mention a temporal expression referring to a fresh date that is equal to the publication date. When the delay $\Delta(d, E)$ increase, the freshness score tends to decrease until it reaches 0.
- Note that when the considered part of document d does not contain any temporal expression close to the entity E, its Freshness score is set to 0.
- The delay between $Date_p(d)$ and $Date_t(x, d)$ is measured in number of days.

As a candidate document may be fresh but not relevant to the entity E, we evaluate a relevance score in order to prioritize fresh relevant documents. *Relevance* score is evaluated as follows:

$$Relevance(d, E) = \prod_{t \in top_k(P_E)} P(t|\theta_d)^{P(t|\theta_{P_E})} \quad (3)$$

P_E is a known relevant page for the entity (for example, the Wikipedia page of the entity).

$top_k(P_E)$ is the set of top k frequent terms in P_E. It can be determined experimentally.

$P(t|\theta_d)$ and $P(t|\theta_{P_E})$ are estimated using a Dirichlet Smoothing as described in Eq. 4

$$P(t|\theta_d) = \frac{tf(t, d) + \mu \frac{tf(t,C)}{\sum_{t' \in C} tf(t',C)}}{|d| + \mu} \quad (4)$$

$tf(t, d)$ is the term frequency of term t in document d.

$tf(t, C)$ is the term frequency of term t in collection C.

C is the reference collection composed from early stream documents before t_0.

μ is a smoothing parameter used to avoid null probabilities.

Finally, we evaluate the vitality score of a document (eq. 5) as the product combination of the *relevance* and *freshness* scores with a $+\epsilon$ smoothing for the freshness score to avoid a null vitality score when the freshness score is equal to zero caused by the absence of temporal expressions.

$$Vitality(d, E) = Relevance(d, E) * (Freshness(d, E) + \epsilon) \quad (5)$$

4. EXPERIMENTS

4.1 Collection

We evaluated our approach within the 2013 CCR task of the TREC KBA track. The task is defined as follows: given an entity identified by a URI (Twitter/Wikipedia), a CCR system strives to recommend to the contributors of a knowledge base relevant documents that are worth citing in a profile of the entity (e.g. its Wikipedia article). The stream corpus contains more than 500 million web documents from several sources (News, Social, Forum, Blog, etc.). It has a size of 4.5 Tera Bytes (compressed), and documents were published in the time range of October 2011 through February 2013. The stream corpus is divided into a training part and an evaluation part. In the latter, we conducted our experiments with **122** entities (persons, organizations, locations). In Table 1, we present some statistics about the KBA 2013 collection. We note that among the 122 entities, only 108 have at least one vital document. Moreover, the higher value of the mean compared to the median indicates that there are some entities that have many more vital documents than others.

Table 1: Some statistics about the KBA 2013 collection

	Training	Evaluation
Time range	Oct.2011 - Feb.2012	Mar.2012 - Feb.2013
Entities with vital(s)	88	108
Total vital	1619	3922
Median of vital	2	6
Mean of vital	13	32

Document annotations were done by the KBA organizers. A document is considered as *vital* if it contains a timely relevant information about the entity (not known before the reference date $t_0 = January,\ 2012$), *useful* (*old-relevant*) if it contains relevant but not timely information about the entity, and *non-relevant* otherwise. We recall that in this work we are interested only in *vital* documents. The official metric for the KBA CCR task is the ***hF1***, i.e. the maximum macro-averaged *F1* measure (Eq. 6). $F1$ is evaluated for each confidence cutoff $i \in]0, 1000]$ (Eq. 7), where 1000 corresponds to the highest level of confidence and 1 corresponds to the level in which all documents are kept.

$$hF1 = Max_i(F1@i) \tag{6}$$

$$F1@i = \frac{2 * mPrecision@i * mRecall@i}{mPrecision@i \ + \ mRecall@i} \tag{7}$$

$$mPrecision@i = \frac{1}{n} * \sum_{E \in \Omega} Precision@i(E) \tag{8}$$

$$mRecall@i = \frac{1}{n} * \sum_{E \in \Omega} Recall@i(E) \tag{9}$$

Ω: set of all evaluated entities

n: number of evaluated entities

$Precision@i(E)$: Precision of E at the confidence cutoff i

$Recall@i(E)$: Recall of E at the confidence cutoff i

4.2 Our proposed method for the CCR task

As indicated in section 3, an entity-vital document filtering system involves two main steps, *filtering* then *scoring*.

Filtering step:.

To reduce the number of documents that are more likely to be non-vital for the entity, we define two sub-steps :

- **Entity Matching** (***E-Matching***):
  A vital document should mention the target entity. As an entity can be mentioned in a document with different variants (surface forms), we collect for each entity a list of variants. For a Wikipedia entity, we use its Wikipedia page title and the bold texts in the first paragraph as variants [6]. For a Twitter entity, we use the display name in the Twitter page as variant. In this level, we retain all documents matching at least one variant of the entity. We alleviate queries in order to capture the maximum number of potential relevant documents. For example, for the Wikipedia entity *Phyllis Lambert*, we can capture documents mentioning the entity by *Phyllis Lambert*, *Phyllis Barbara Lambert* or *Phyllis B Lambert*, etc.

- **Filtering spam documents** (***Filters***):
  Excluding spam documents could improve system performance. Therefore, we define three filters in order to reject documents matching the entity but more likely to be spam:

  - *A language filter* that removes all documents recognized as Non-English-documents using a Java language detector[1].
  - *An enumeration filter* that removes documents that mention the entity only in an abusive list of more than n entities. We set n to 30 based on some observations in the training time range.
  - *A link filter* that removes documents that contain more than *20* hyper-links (based on some observations in the training time range).

Scoring step:.

We evaluate the vitality score of documents that passed the filtering step based on *topicality and freshness* as described in Eq. 1. *Freshness* is reflected by the dates recognized from temporal expressions contained in the document text. Temporal expressions such as *Last week*, *At the end of 2012*, *December 31, 2013*, etc. are identified and normalized using SUTIME [5]. This library uses the document publication date as reference. For example, for a document from 2014-01-01, SUTIME would resolve the date referred to by *Tuesday* as 2013-12-31.

We considered three configurations of our approach:

- **T&F$_{\mathbf{Sen}}$**: The freshness is estimated considering the temporal expressions recognised from the *sentences* mentioning the entity.
- **T&F$_{\mathbf{Pg}}$**: The freshness is estimated considering the temporal expressions recognised from the *paragraphs* mentioning the entity.
- **T&F$_{\mathbf{W}}$**: The freshness is estimated considering the temporal expressions recognised from the *whole document*.

To evaluate the impact of *freshness*, we run another configuration that considers only the *topicality* of documents (Eq. 3); we denote it by **onlyT**.

[1] www.jroller.com/melix/entry/nlp_in_java_a_language

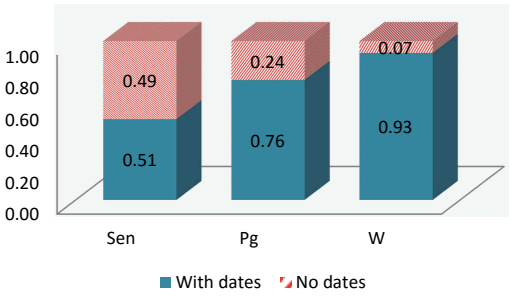

Figure 2: Availability of dates in the different parts of all documents that pass the filtering step

The above configurations are evaluated with and without using spam filters. When spam filters are used, we add +**S** to the configuration label.

To estimate the topical relevance of the document (Eq. 3), we use a known relevant page (P_E) to get information about the entity. For a Wikipedia entity, P_E consists of the Wikipedia article of the entity (on January 1st, 2012), and for a Twitter entity, P_E contains only the display name in the entity Twitter page.

Our approach uses 3 parameters σ, μ and $top_k(P_E)$. To estimate their optimal values, we use a 3-fold cross-validation method. We vary $\sigma \in [1, 360]$ (step=30), $\mu \in [50, 1000]$ (step=50) and $top_k(P_E) \in [5, 30]$ (step=5). Optimal values are: $\sigma = 30$, $\mu = 200$ and $top_k(P_E) = 20$. We set ϵ in eq. 5 to 10^{-4}.

The CCR task requires systems to assign a confidence score $\in [1, 1000]$ to each document. We adopted the following strategy; rank 1 gets a confidence value 1000, rank 2 gets a confidence value 999, etc.

4.3 Results

4.3.1 *Recognized dates in documents*

The freshness score calculated in equation 1 is based on the recognition of temporal expressions mentioning the entity. However, in some cases, we can fail to detect a temporal expression in one or some parts (Sen, Pg, W) of the document. Figure 2 shows the availability of dates in the different parts of all documents matching the studied entities.

We can see that most of the documents (93%) contain at least one temporal expression in their bodies. Considering only smaller parts that mention the entity, we can recognize at least one date in a large part of paragraphs (76%) and in the half of sentences.

In our approach, we hope that in vital documents we can find more fresh dates than in other documents classes which means that the optimal temporal distance in vital documents ($\Delta(d, E)$) are expected to be low. We further analyse documents by classifying $\Delta(d, E)$ into two ranges : "one-year-fresh" range when the distance is less than one year and "old" otherwise. Figures 3, 4 and 5 plot the results by distinguishing the three different classes : vital, useful and non-relevant respectively.

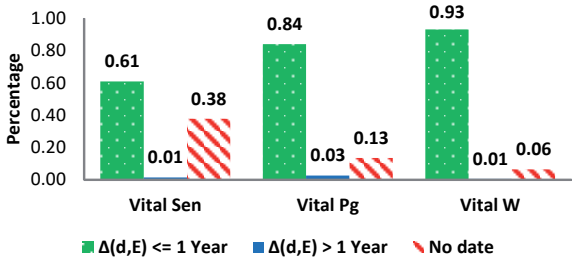

Figure 3: Freshness of dates in vital documents

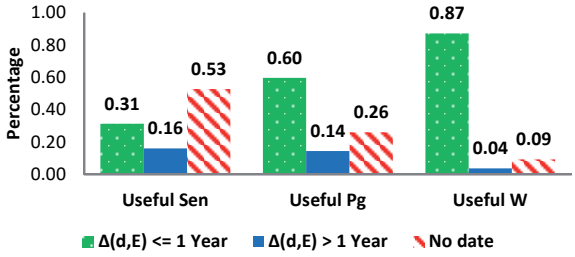

Figure 4: Freshness of dates in useful documents

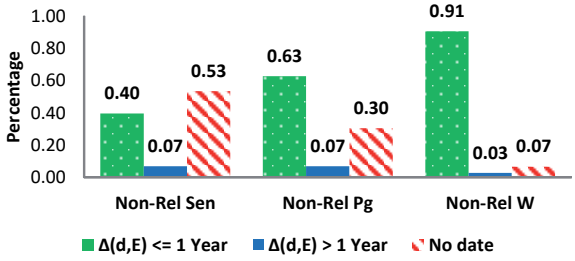

Figure 5: Freshness of dates in non-relevant documents

We can see that the percentage of one-year-fresh dates are greater in vital documents than in useful or non-relevant documents. The difference is notable, especially when focusing on the sentences or the paragraphs. 84% of vital paragraphs contain a temporal expression referring to a fresh date comparing to the publication date of the document (i.e., $\Delta(d, E) < 1year$). In addition, we notice also that old dates are more likely to be mentioned in useful documents that describe old-relevant information about the entity.

In the following section we compare our system configurations that leverage temporal expressions recognized in the different parts of documents.

Table 2: Comparison of the different configurations of our approach. i* corresponds to the confidence cut-off in which F1 is maximum.

	i*	mPrec.@i*	mRec.@i*	hF1
$T\&F_{Sen}+S$	910	**0.281**	0.637	0.390
$T\&F_{Pg}+S$	910	**0.281**	0.671	**0.396**
$T\&F_{W}+S$	370	0.249	**0.783**	0.378
$onlyT+S$	340	0.248	**0.783**	0.377
$T\&F_{Sen}$	590	0.229	0.765	0.352
$T\&F_{Pg}$	920	0.256	0.617	0.362
$T\&F_{W}$	880	0.232	0.673	0.345
$onlyT$	870	0.229	0.675	0.342

4.3.2 Comparison of the different configurations

Table 2 compares the different configurations of our approach. We observe that exploiting the freshness and topicality ($T\&F_*$) improves results compared to using only the topicality without leveraging temporal expressions ($onlyT$). This confirms that freshness represents an important factor to detect vital documents. Estimating freshness using only parts of text describing the entity (Sen or Pg) performs better than using the whole document content. One reason could be that considering some parts that are not in the proximity of references to the entity could bring fresh dates (closer to the publication date) which are not related to the target entity. Searching temporal expressions only in sentences mentioning the entity ($T\&F_{Sen}*$) may be insufficient in some cases to detect dates related to the entity (as shown in figure 3, 61% of sentences mentioning the entity in vital documents contain a fresh date), which can explain the slight improvement when considering the paragraphs mentioning the entity ($T\&F_{Pg}*$) where fresh dates are present in 84% of them. The high values of the optimal confidence cutoff ($i^* = 910$) indicates that $T\&F_{Sen}+S$ and $T\&F_{Pg}+S$ rank well vital documents.

We also notice that all configurations perform better when applying spam filters. Figure 6 shows the impact of each of the filters in hF1 when added to $T\&F_{Pg}$. We can observe that all of them have a good precision (97%) which means that they do not reject many vital documents. $hF1$ is well improved especially when using *link* and *enumeration* filters which indicates that the test collection contains many documents mentioning the target entity in a spam way without providing any relevant information about it. Using all filters rejects many non-vital documents (about one-fourth) which improves the filtering step performance in terms of $hF1$ (+0.032). This observation supports the hypothesis made in [8]: *well-crafted Boolean queries can be effective filters for CCR task.*

4.3.3 Vitality probability given the value and the position of the optimal date

In figure 7 we analyse the probability of document d be vital given the value of the optimal delay $\Delta(d, E)$ and the considered part in the document (Sen, Pg or W).

The first remark worth making is the presence of a fresh

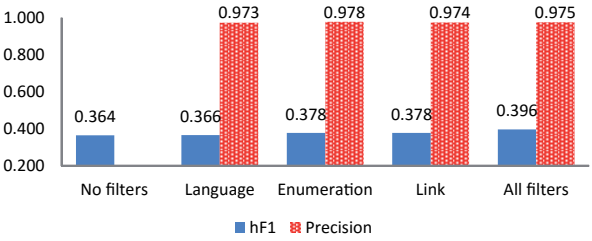

Figure 6: Impact of each spam filter on hF1 when added to $T\&F_{Pg}$. $Precision = \frac{\#non\ vital\ rejected\ docs}{\#all\ rejected\ docs}$

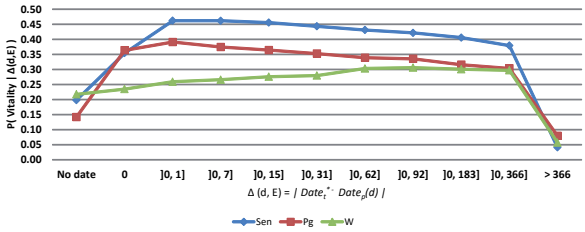

Figure 7: Vitality probability of a document given of the optimal delay and the considered part in the document (Sen, Pg, W)

date (what ever its value) in the sentence mentioning the entity gives a better indication of the document vitality (up to 46%) than its presence in a farther position (i.e., Pg or W). In addition, the fresher this date is, the higher the vitality probability is, except when the optimal delay $\Delta(d, E)$ is zero. The reason of this exception is that many non-vital documents mention their publication date in the beginning or in the end of the body. We can also notice that when $\Delta(d, E)$ is greater that one year, the vitality probability of the document becomes very low. This observation can be helpful to discard many non-vital documents from stream. Finally, when no date is recognized in the document, the probability of vitality is low but not too low, which requires to exploit other features to decide whether the document is vital or not.

4.3.4 Our approach vs. state-of-the-art approaches

In this section, we compare our best configuration ($T\&F_{Pg}+S$)[2] with the top 3 CCR systems: **BIT**[14], **Umass**[7] and **Udel**[13].

Table 3 compares system performances in the *filtering* step considering all returned documents (i.e., *confidence cutoff =1*). We denote our approach in the filtering step by **E-Matching** when no filter is used and **E-Matching+filters** when filters are applied. *Ranking strategies* (our approach, Umass, Udel) perform generally better in recall than the *classification strategy* (BIT) that can be penalized if it fails to properly classify many vital documents. Our matching

[2]Comparison is done using the official scorer of the task. Our best run is available at this URL: http://www.irit.fr/~Rafik.Abbes/SAC15Runs/

Table 3: Comparison of our approach with 2013 CCR systems in the filtering step.

	mPrec.@1	mRec@1	F1@1
Udel	0.199	0.695	0.309
Umass	0.201	0.662	0.309
BIT	0.244	0.650	0.355
E-Matching	0.217	**0.794**	0.354
E-Matching + Filters	**0.248**	0.782	**0.377**

method performs best among the proposed methods in terms of recall which can be explained by the use of alleviated queries in the entity matching sub-step.

Table 4 compares systems using the official metric of the task (hF1). Our approach outperforms the best proposed system in the task (BIT). Significance test is not suitable in the case of $mPrecision@i^*$, $mRecall@i^*$ and $hF1@i^*$ because the optimal confidence cutoff i^* and the confidence scoring strategy are not the same for each system.

Table 4: Comparison of our approach with the 2013 CCR systems. i* corresponds to the confidence cutoff in which F1 is maximum.

	i*	mPrec.@i*	mRec.@i*	hF1
Udel	40	0.199	0.695	0.309
Umass	660	0.216	0.591	0.316
BIT	140	0.257	0.601	0.360
$T\&F_{Pg}+S$	910	**0.281**	**0.671**	**0.396**

To evaluate in depth the performance of our system compared to the best proposed system (BIT) in the task, we consider the top-30 returned documents of each system. We choose 30 as it represents the mean of vital documents per entity. Table 5 shows the results using the macro averaged precision ($mPrec.@T30$), recall ($mRec@T30$) and F-measure ($mF1@T30$). Results show that our approach ($T\&F_{Pg}+S$) significantly improves the ranking of vital documents compared to BIT. Figure 8 illustrates the difference in $F1@T30$ of each entity between $T\&F_{Pg}+S$ and BIT. We can observe that $T\&F_{Pg}+S$ performs better than BIT in 69 topics and worse in 25.

Table 5: Comparison of our best configuration ($T\&F_{Pg}+S$) with the best proposed system (BIT) in the task . † denotes a significant improvement (we used the paired t-test with p <0.05). *@T30* means considering the top-30 documents

	mPrec.@T30	mRec.@T30	mF1@T30
BIT	0.211	0.297	0.170
$T\&F_{Pg}+S$	**0.286** †	**0.489** †	**0.262** †

Let us take for example the entity *Hoboken Volunteer Ambulance Corps*. This topic has 32 vital documents in the evaluation time range. Regardless of the ranking (considering

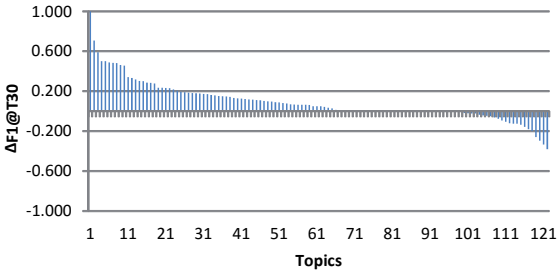

Figure 8: Difference in $F1@T30$ between $T\&F_{Pg}+S$ and BIT for each topic

all the returned documents, i.e., *confidence cutoff=1*), our approach ($T\&F_{Pg}+S$) returns 66 documents containing 31 vital ones, whereas BIT returns 117 documents containing 29 vital ones. We can see that both methods perform well in recall (> 90%). However, our method rejects more non-vital documents than BIT due to the use of spam filters which allow rejecting many spam documents such as the third document in Figure 9. In terms of ranking, considering the top-30 returned documents of each system, BIT gets only one vital document, whereas, our approach ($T\&F_{Pg}+S$) detects 16 vital documents which corresponds to 50% of recall. These documents contain fresh dates in the proximity of the entity (such as the first two documents in Figure 9), which can explain the good performance.

4.3.5 Real cases from the TREC KBA 2013 corpus

Table 6 shows some real case examples from the TREC KBA 2013 corpus. In the first part, we show examples in which temporal expressions recognized either implicitly (line 1) or explicitly (line 2) from the documents are helpful to infer their vitality as the delay of the inferred dates to the publication dates are low. The large delay in the third example (line 3) infer correctly the right class of the document.

In the second part of the table, we give some examples in which using only the temporal expressions is insufficient for many possible reasons :

- some vital documents do not mention a date near to the entity (lines 4 and 5),
- the delay is high because the vital document mention an upcoming event in the far future (line 6),
- the normalisation of the recognized date is wrong like in line 7 where the expression *'to this day'* refers to a past fact whereas the tool that we used to recognize dates considered that this expression refers to the present and therefore the delay is estimated to 0,
- some non-relevant documents (lines 8, 9 and 10) can mention a new date, so the freshness score should be combined by a relevance score. We remark that the relevance score is better estimated for wikipedia entities for which we have some context information (the wikipedia page of the entity), whereas for twitter entities only the entity name is used.

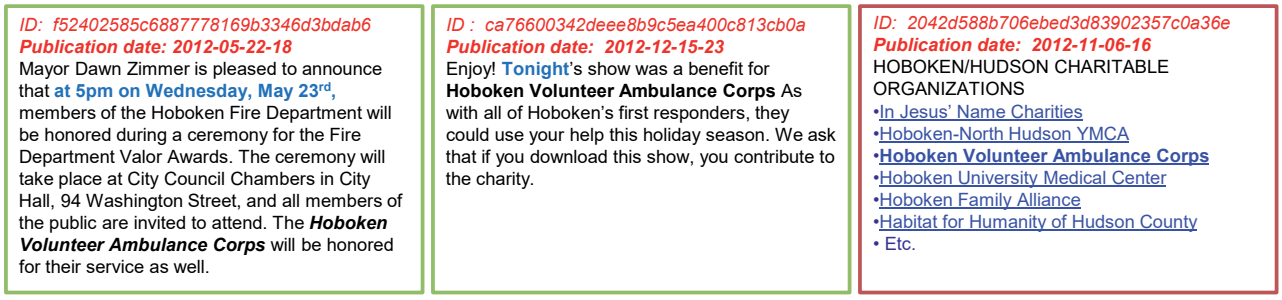

Figure 9: Examples of documents for the topic *Hoboken Volunteer Ambulance Corps*, the first two documents are vital and the last one is not relevant rejected by our spam filters

Table 6: Real cases from the TREC KBA 2013 corpus when the temporal expressions are helpful and when they are insufficient

#	Entity (E)	Sentence (Sen)	$Date_p(d)$	$\Delta(d, E)$	class
		cases where temporal expressions are helpful			
1	Atacocha	On ***Friday***, silver miner Minera *Atacocha*, the Lima-based zinc and silver mining company, fell by 5.4%	2012-03-02-04	1 day	vital
2	Barbara Liskov	***Monday, 05 March 2012*** *Barbara Liskov* is among the 2012 inductees to the National Inventors Hall of Fame in recognition of her contributions to programming languages and system design.	2012-03-05-12	0 day	vital
3	Brenda Weiler	The local chapter's community walk was started in 2006 by *Brenda Weiler*, of Fargo, after she lost her older sister to suicide ***the year prior***.	2012-09-22-05	265	old-relevant
		cases where temporal expressions are insufficient			
4	Angelo Savoldi	The NWA is pleased and proud to induct *Angelo Savoldi* into the Hall of Fame!	2012-11-14-13	No date!	vital
5	Barbara Liskov	Murray on mathematical biology, *Barbara Liskov* of MIT on in modern programming languages, Ronald Rivest of MIT on cryptography, Leslie G.	2012-04-30-20	No date!	vital
6	evvnt	The 13th Annual European Shared Services & Outsourcing Week The 13th Annual European Shared Services & Outsourcing Week The 13th Annual European Shared Services & Outsourcing Week offered by *evvnt* will take place in Prague on ***21 May 2013***	2012-11-14-06	188 days	vital
7	Bob Bert	Drummer *Bob Bert* played on the album but departed before the tour, and his replacement Steve Shelley remains to ***this day***.	2012-11-08-12	0 day	old-relevant
8	Tony Gray	Port continued their pre-season schedule with a 4-0 win at Radcliffe Borough on ***Tuesday night*** with goals from Steven Tames (two), Shaun Whalley and *Tony Gray*	2012-08-02-09	0 day	non-relevant
9	Alexandra Hamilton	By *Alexandra Hamilton* Email the author ***March 6, 2012*** Tweet Email Print 1 Comment ? Back to Article new Embed | Share	2012-03-06-12	0 day	non-relevant
10	Blair Thoreson	*Blair Thoreson* has served in the North Dakota House of Representatives since 1998, representing District 44. Tags: ALEC, American Legislative Economic Council, *Blair Thoreson*, van jones This entry was posted on Wednesday, ***April 18th, 2012 at 12:19 pm*** and is filed under Blog	2012-04-18-19	0.5 day	non-relevant

5. CONCLUSION AND FUTURE WORK

In this paper, we are interested in filtering vital documents related to an entity from a document stream. We propose an approach that evaluates the freshness of temporal expressions that mention the entity with regard to the publication date of the document to infer the vitality. Experiments carried out over the 2013 TREC KBA collection confirm the usefulness of leveraging temporal expressions to detect vital documents. We gave some real examples in which leveraging temporal expressions can be helpful is some cases or insufficient in others. The absence of temporal expressions in some documents, or in a specific considered part like the sentences mentioning the entity, requires to exploit other factors in order to estimate the vitality. In addition, we show that applying some boolean filters leads to substantial improvement in the system's performance. Further work will concern a way to combine the freshness of temporal expressions with other factors that can infer vitality such as detecting bursts in the stream or exploiting some action patterns. In addition, we would like to investigate how to automatically extract vital information from the detected vital documents. In this context, we made a preliminary work described in [1] that extracts only the interesting sentences from the document stream. In the short term, we intend to extend our system by using knowledge extraction tools in order to automatically update knowledge bases.

6. REFERENCES

[1] R. Abbes, K. Pinel-Sauvagnat, N. Hernandez, and M. Boughanem. Accelerating the update of knowledge base instances by detecting vital information from a document stream. In *2015 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Singapore, December 06-09, 2015*, 2015. (to appear).

[2] K. Balog and H. Ramampiaro. Cumulative citation recommendation: Classification vs. ranking. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 941–944, New York, NY, USA, 2013. ACM.

[3] K. Balog, H. Ramampiaro, N. Takhirov, and K. Nørvåg. Multi-step classification approaches to cumulative citation recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 121–128, Paris, France, 2013.

[4] L. Bonnefoy, V. Bouvier, and P. Bellot. A weakly-supervised detection of entity central documents in a stream. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 769–772, New York, NY, USA, 2013. ACM.

[5] A. X. Chang and C. Manning. Sutime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[6] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the Joint Conference on EMNLP-CoNLL*, EMNLP-CoNLL'07, pages 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[7] L. Dietz and J. Dalton. Umass at TREC 2013 knowledge base acceleration track: Bi-directional entity linking and time-aware evaluation. In *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, 2013.

[8] M. Efron, C. Willis, and G. Sherman. Learning sufficient queries for entity filtering. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1091–1094, New York, NY, USA, 2014. ACM.

[9] J. R. Frank, S. J. Bauer, M. Kleiman-Weiner, D. A. Roberts, N. Tripuraneni, C. Zhang, C. Ré, E. M. Voorhees, and I. Soboroff. Evaluating stream filtering for entity profile updates for TREC 2013. In *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, 2013.

[10] J. R. Frank, M. Kleiman-Weiner, D. A. Roberts, F. Niu, C. Zhang, C. Ré, and I. Soboroff. Building an entity-centric stream filtering test collection for TREC 2012. In *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012*, 2012.

[11] J. R. Frank, M. Kleiman-Weiner, D. A. Roberts, E. M. Voorhees, and I. Soboroff. Evaluating stream filtering for entity profile updates in TREC 2012, 2013, and 2014. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*, 2014.

[12] X. Li, C. Li, and C. Yu. Entity-relationship queries over wikipedia. *ACM Transactions on Intelligent Systems and Technology*, pages 70:1–70:20, 2012.

[13] X. Liu, H. Fang, and J. Darko. A related entity based approach for knowledge base acceleration. In *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, TREC'13, Gaithersburg, USA, 2013.

[14] J. Wang, D. Song, C.-Y. Lin, and L. Liao. BIT and MSRA at TREC KBA CCR Track 2013. In *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, TREC'13, Gaithersburg, USA, 2013.

[15] M. Zhou and K. C.-C. Chang. Entity-centric document filtering: Boosting feature mapping through meta-features. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 119–128, New York, NY, USA, 2013. ACM.